Automated Diagnosis of Breast Cancer Using Artificial Intelligence

Through the Implementation of Neural Networks
An Open Source, Cloud Implemented Diagnosis

Guss, William
carlostnb@gmail.com

Chen, Patrick
patrickbochen@gmail.com

March 11, 2014

Abstract

The challenge of diagnosing cancer is that no single test can accurately succeed. Diagnostic testing is essential to evaluate the health of an individual and determine whether the individual has cancer. Diagnostic imaging is a useful technique to produce an internal picture of the body for analyzing structure. However, medical professionals are required to successful analyze the images and determine whether the individual has cancer. Applying artificial neural networks to this problem makes the analysis more efficient while minimizing error in diagnosis.

The purpose of the project is to implement a successful neural network with backpropagation to analyze a breast cancer numerical and image dataset. It also evaluates the efficiency of the network as it is influenced by different conditions. The efficiency is gauged by the error percentages accumulated by the network. Furthermore, statistical analysis is applied to the network in order to analyze the effectiveness.

The project showed that despite the adaptability of the neural network, it is still unable to remove the error completely. While neural networks are useful, they cannot be relied on completely. However, there is a tradeoff between error and flexibility in the network. While the error has the potential to be removed from the testing of the neural network, the network would be over fitted to the dataset it is being trained and tested on so that it would lose its ability to successfully analyze similar forms of data.

Contents

1 Introduction

Breast cancer has become one of the most common existing cancers. It is a frequent and leading cause of mortality, especially in developed countries. This risk of receiving the cancer also increases with age. However, early detection of breast cancer leads to increases in survival rates. Moreover, doctors may employ several methods to determine the existence of cancer. One diagnostic procedure that investigates potential lumps in the breast is called a fine-needle aspiration biopsy (FNA). These biopsies are very safe and minor surgical procedures making them viable options for wide usage for diagnosis. A thin hollow needle is inserted into the breast to sample cells, which are then stained and examined under a microscope. A pathologist or another expert that examines the biopsy will then determine the state of the breast based on observations made from the data. However, due to the fact that the final decision is subjectively made by the pathologist, there is still large room for human error. Thus implementing artificial intelligence, specifically neural networks, is a more accurate and efficient way of processing the data and determining the malignancy of the breast.

Despite these options, the most current diagnostic method for early detection of breast cancer is a mammography. These are x-rays of the breasts that detect micro-calcifications. Micro-calcifications are tiny bits of calcium that usually indicate extra cell activity in the breast tissue which can indicate early breast cancer. However the majority of scattered micro-calcifications, which appear as white speckles on a mammography, are benign. There micro-calcifications range from one hundred micrometers to two millimeters. Because of the difficulty in locating and analyzing these aspects and because survival from breast cancer is dependent on early detection, a computer aided diagnosis is useful and beneficial to detect micro-calcification clusters.

Because, these micro-calcifications represent such a large range of possible image datasets, a complex function must be used to model the existence of these in relation to the benign or malignant state of the cells in the breast. A neural network is more optimal for determining this function as opposed to normal algorithmic approaches that are unable to model the function as well. Because increasing the survival rate from breast cancer is dependent on correctly identifying the cancer to begin with, using a more effective method with low error rates such as a neural network is preferred.

2 Background

2.1 Neural Networks Compared to Conventional Approaches

Neural networks take a different approach to problem solving than conventional algorithmic approaches. Algorithmic problem solving requires a fixed set of actions to determine a solution, and if absent, an algorithmic function for such a problem is impossible, restricting the problem solving capabilities of conventional computing. Neural networks, in contrast, learn by example and cannot be programmed to perform a specific task, allowing a computer system to approximate an otherwise unknown function. Neural networks, as a result, are able to perform adaptive learning, self-organization, real time operation, as well as fault tolerance. The reliability of a neural network fails to match that of a conventional algorithm, as operation under certain conditions can be unpredictable.

Often, to combat uncertainties in the neural network and the limitations of an algorithmic approach, the two problem solving methods are combined to perform a task at high efficiency.

2.2 Similarities Between Neural Networks and Humans

Neural networks are a form of artificial intelligence that derives from the structure of biological nervous systems and data processing methods within those systems. Both the brain and neural networks are composed of a large number of processing elements, referred to as neurons, which are interconnected by weights (in artificial neural networks) or axons and dendrites (in biological neural networks), which work together with the neurons to solve problems. Artificial networks and an organism's nervous system learn by example. These two structures update their weights or connections in response to inputs and whatever result is desired.

Whereas artificial networks are typically data intensive and thus limited to several hundred units, biological neurons can consist of 10,000 individual inputs, immensely more complex. The complexity of the existing neural network is limited by the computing power of the computers or artificial systems in place today.

2.3 Neural Networks

To account for the lack of an algorithm, neural networks attempt to discern patterns from a dataset. Although any dataset is able to work, a larger dataset is typically required to allow a more adaptive nature for the neural network. The neural network takes in a set of inputs and passes on the values through a series of numbers termed weights, which then pass on an altered value to neurons in subsequent layers (either hidden or output). Once the network outputs values, the results are compared to the desired output, and the network's weights are adjusted accordingly, through the method of backpropagation.

The feed forward neural network is the simplest type of an artificial neural network utilized. Information moves in a single direction, forward, from the input nodes, through hidden nodes (if any) and to output nodes. There are no cycles or loops in the network. Feed forward networks are the most popular and widely used function-modeling structures that reflect a dataset.

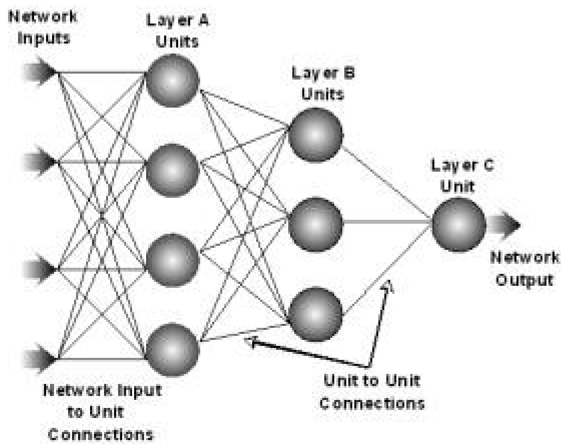

Figure 1: A diagram of a feed-forward artificial neural network.

2.3.1 Advantages

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.

2. Self-Organization: A neural network can create its own organization or representation of the information it receives during learning time.

3. Real Time Operation: Neural network computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.

4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage

2.3.2 Limitations

1. Limited by the computing power of existing computers or current artificial systems.

2. Unpredictable at times

3. Randomness due to the initial starting weight values which are determined and randomly based off a Gaussian distribution curve. At times, the randomness leads to the inability of the network to converge.

4. Unable to completely remove error because of the necessity of some in order to retain the adaptive nature of the network.

5. Extremely hard to optimize and minimize the error because of the numerous amount of variables that make up the network and affect it.

2.4 Error Backpropagation

Backpropagation, or backward propagation of error, is the most common algorithm for training neural networks. It is used to find a function that best models the input data. The goal of the backpropagation algorithm is to train any neural network such that it can learn to create any arbitrary map from inputs to outputs. From the error between the desired output and the actual output, the network readjusts its weights to reduce the error being produced. Backpropagation is most useful for feed forward networks which have no feedback.

2.4.1 Intuition

Each neuron employs a linear output that is the weighted sum of its input. Initially, before training, the weights will be set randomly. Afterwards, the neuron learns from training input by employing the sum squared error to measure the discrepancy between the expected output and the actual output. Therefore, the problem of mapping inputs to outputs can be reduced to an optimization problem of finding a function that will produce the minimal error. The backpropagation algorithm aims to find the set of weights that minimizes the error. The method used in backpropagation, to find the minima of a function in any dimension that models that map from input to output which minimizes the error, is gradient descent.

2.4.2 Derivation

Error backpropagation is derived through a series of chain rule calculations based on the sum squared error function of the neural network. The error function is composed of a series of output vectors which themselves are composed of weight vectors and output vectors from activation functions of neurons from previous layers. The essential goal of this derivation is to find a suitable weight update rule such that the error function of the neural network descends to a local minimum using a weight gradient vector.

Take the sum squared error function,

$$E = \frac{1}{2}\sum_{j\in L}(d_j - \sigma_j)^2$$

where L is the output layer and j is a given node on the output layer. To calculate $\frac{\partial E}{\partial w_{ij}^L}$, chain rule is direct given a lack of composite pathways to that weight. Therefore the partial derivative is given by

$$\frac{\partial E}{\partial w_{ij}^L} = \frac{\partial E}{\partial \sigma_j^L}\frac{\partial \sigma_j^L}{\partial \text{net}^L}\frac{\partial \text{net}^L}{\partial w_{ij}^L}$$

Given that $\sigma = \tanh(\text{net})$ and $\text{net} = \sum w_{ij}\sigma_i$, we calculate the (i, j) component of the weight gradient on the output layer to be

$$\frac{\partial E}{\partial w_{ij}^L} = (d_j - \sigma_j^L)\sigma_j^L(1 - \sigma_j^L)\sigma_i^{L-1}$$

Although this calculation of weight gradient is applicable to the last layer of weights, a generalization must be made using higher-order chain rule to determine the weight gradient for a weight on any given layer, l. For the sake of simplicity we will define a node delta for the output layer as

$$\delta_j^L = \frac{\partial E}{\partial \sigma_j^L} \frac{\partial \sigma_j^L}{\partial \mathrm{net}^L}$$

Thereafter follow that each weight gradient is directly related to every pather leading to the weight, itself. Therefore for the hidden layer, $l = L - 1$, we must define the node delta as a summation given by

$$\delta_i^{L-1} = \sum_{j \in L)} \frac{\partial E}{\partial \sigma_j} \frac{\partial \sigma_j}{\partial \mathrm{net}_j} \frac{\partial \mathrm{net}_j}{\partial \sigma_i} \frac{\partial \sigma_i}{\partial \mathrm{net}_i}$$

Using the gneralization employed in deriving the node delta for a given neuraon in the first hidden layer, we can then state the generalization for any hidden layer to be

$$\delta_i^l = \sum_{j \in (l+1)} \delta_j w_{ij} \frac{\partial \sigma_i}{\partial \mathrm{net}_i}$$

Now that we have a model for calculating the node delta for any given neuron in any given layer, let

$$\frac{\partial E}{\partial w_{ij}^l} = \delta_j^{l+1} \sigma_i^l$$

To calculate the change in weight, error backpropagation then uses the method of gradient descent to define the weight update as

$$\Delta w(t) = -\lambda \frac{\partial E}{\partial w_{ij}^l} + \mu \Delta w(t-1)$$

where λ is a user-defined learning rate and μ is a momentum term which avoids local minima trapping.

2.5 Haar Wavelet Transform

Haar wavelet is a subset of a larger concept known as wavelet transform. Wavelet transformations are meant to change the time-frequency of an image. The transformation however only occurs in the time extension while maintaining the shape. An application of wavelet transformation is to compress images while reducing loss of data and is used instead of the Fourier transformation because it is able to capture both frequency and location information. Inside wavelet transforms, there are two sub-categories, discrete and continuous. Discrete is represented by integers while continuous can be represented over an entire range of numbers. Discrete is preferable for image recognition because it is used to represent pixel values which are whole numbers. These pixel values once altered by wavelet transform are then used as input for a neural network.

The exact wavelet transform used during this experiment was Haar wavelet transform. It is a transform that creates a sequence of rescaled rectangular shaped functions. For the purpose of this experiment, the transform is meant to reduce the data that is being inputted into the neural network to increase processing speed. It does so by using minimal pixels and using multiple scales which are represented simultaneously. Other advantages of the process are that it is both invertible and linear.

2.5.1 Decomposition

Given an original image with pixel values P_1,P_2,P_3, and P_4 as represented in diagram below and a size length of 2^n where n is an arbitrary integer.

P_1	P_2	...	...
P_3	P_4	...	...
...	...	...	...
...	...	...	...

After one wavelet transform, the pixel values of the new image are represented by

$(P_1 + P_2 + P_3 + P_4)/4$	...	$(P_1 - P_2 + P_3 - P_4)/4$	...
...	...	...	...
$(P_1 + P_2 - P_3 - P_4)/4$	...	$(P_1 - P_2 - P_3 + P_4)/4$	...
...	...	...	...

The two basic operations that occur are the sum of the four pixels and the difference of pair wise sums of pixels to create a new image. This process continues for part of the image, specifically for the length range of 0 to 2^{n-1}. This process will continue recursively until the length range reaches two. It is meant to reduce the amount of input to the neural network while keeping the amount of information constant.

3 Datasets

3.1 Breast Cancer Wisconsin (Original) Dataset P-FNA Test

This dataset is numerical and multivariate with integer attribute values. The creator was Dr. William H. Wolberg at the University of Wisconsin in Madison, Wisconsin. The diagnostic test created for this specific dataset is called a P-FNA test, proportional fine needle aspiration test.

The dataset used had 9 input attributes, each from a range of 1 to 10. There were a total of 699 data points. However, 16 of the points had inconsistencies where a question mark stood in place of a number. The 16 data points thus were excluded from both the training and the testing of the neural network. 10% of the data was used for testing while the other 90% was used for training.

While the data had the output of 2 for benign and 4 for malignant, during the testing of the network these numbers were changed to 0 and 1 for benign and malignant respectively. This is because the logistic function can only output from a range of -1 to 1. For the actual dataset, 65.5% were benign and the other 35.5% were malignant. The source also claimed that there is also a 5% discrepancy in the dataset.

3.1.1 Inputs

Clump Thickness
: Benign cells tend to be group in a monolayer, while cancerous cells are often grouped in a multilayer.

Uniformity of Cell Size
: Cancer cells tend to vary drastically in size and shape, thus a lower uniformity correlates with a higher possibility of cancer cells.

Uniformity of Cell Shape
: Cancer cells tend to vary drastically in size and shape, thus a lower uniformity correlates with a higher possibility of cancer cells.

Marginal Adhesion
: Cancer cells tend not to stick to one another as well as normal cells, so less adhesion correlates to a higher malignancy.

Single Epithelial Cell Size
: The size is related to uniformity, but enlarged epithelial cells may be malignant.

Bare Nuclei
: It is an index of nuclei not surrounded by a cell, which is present in malignant tumors.

Bland Chromatin
: Uniformity of texture appears in benign cells, while malignant tumors are typically coarse-textured. A lower number corresponds to more unity.

Normal Nucleoli
: The rate of occurrence of normal nucleoli; abnormal nucleoli indicate possible mutated DNA, thus possible genetic expression for cancer reproduction. Thus, the smaller the rate of occurrence, the larger the chance of malignancy becomes.

Mitoses

Cancer cells tend to replicate faster which contributes to a tumor and leads to increased potential in harmful consequences. Thus, a set of cells with a higher rate of mitoses has an increased chance of being malignant.

3.2 Breast Cancer Wisconsin (Diagnostic) Dataset D-FNA Test

This dataset is numerical and multivariate with real attribute values. The creator was Dr. William H. Wolberg at the University of Wisconsin in Madison, Wisconsin. The diagnostic test created for this specific dataset is called a D-FNA test, detailed fine needle aspiration test.

The dataset used had 30 input attributes, each represented by a real value with four significant digits. Ten features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The mean, standard error, and largest of these features were computed for each image, resulting in 30 real valued features. There were a total of 569 data points. 10% of the data was used for testing while the other 90%t were used for training.

While the data had the output of B for benign and M for malignant, during the testing of the network these numbers were changed to 0 and 1 for benign and malignant respectively. This is because output of the neural network is a numerical value. For the actual dataset, 62.7% were benign and the other 37.3% were malignant.

3.2.1 Inputs

Radius

The mean of the distances from the center of the nucleus to points on the perimeter.

Texture

The standard deviation of grey-scale values of the FNA.

Perimeter

The perimeter of the nucleus.

Area

The area of the nucleus.

Smoothness

The local variation in radius lengths.

Compactness

Calculated by $\frac{P^2}{A} - 1$, where P is the perimeter and A is the area.

Concavity

The severity of concave portions on the countour of the nucleus.

Concave points

The number of concave points on the contour.

Symmetry

An average measure of symmetry across the nucleus.

Fractal dimension

The coastline approximation less one is the fractal dimension, which is also noted to be an objective and reproducible measure of the complexity of the tissue architecture of the biopsy specimen.

3.3 Mammographic Image Analysis Society (MIAS) Database

This dataset is composed of images from mammographies. The creator was the Mammographic Image Analysis Society. The diagnostic test created for this specific dataset is called a MID test, mammography image diagnostic test.

The dataset is composed of images with 200 micron pixel edges. Each image size is 1024x1024 pixels. The dataset used for input to the neural network is called a mini-MIAS database created by J. Suckling who reduced the resolution of the original MIAS database. There were a total of 322 data images. 18.2% of the data was used for testing while the other 81.8% were used for training. However, only a small portion of the actual images were used. The breasts that were in the dataset had three distinct background tissues: fatty, fatty-glandular, and dense-glandular. Only images with fatty background tissue were used during the experiment to reduce a variable for the network. Also only images of left breasts were used to reduce the amount of uncertainty in the network because of the existence of black space on either side of the breast. Once these specific images were separated from the original dataset, a total of 22 images remained. Of these 11 were tumorous and 11 were normal (more normal ones existed but in order to achieve consistency during training and testing only 11 were used).

While the data had the output of N for normal and M for malignant, during the testing of the network these numbers were changed to 0 and 1 for normal and malignant respectively. This is because output of the neural network is a numerical value. For the actual dataset, 50% were malignant, and the other 50% were normal.

4 Implementation